PERSONNEL PSYCHOLOGY
2011, 64, 53–87

IS MORE STRUCTURE REALLY BETTER? A COMPARISON OF FRAME-OF-REFERENCE TRAINING AND DESCRIPTIVELY ANCHORED RATING SCALES TO IMPROVE INTERVIEWERS' RATING QUALITY

KLAUS G. MELCHERS
NADJA LIENHARDT
MIRIAM VON AARBURG
MARTIN KLEINMANN
Universität Zürich, Switzerland

This study provides the first comparison of 2 methods proposed to increase the structure of selection interviews: frame-of-reference (FOR) rater training for interviewers and providing interviewers with descriptively anchored rating scales. In contrast to descriptively anchored rating scales, evidence for the efficacy of FOR training for interviewers is still missing even though its effects have been established in other domains. To evaluate the effectiveness of the 2 methods, we used a 2×2 design in which both methods were manipulated independently. Participants observed and rated different interviewees' performance in a set of videotaped interviews. We found that both methods led to substantial, and comparable, improvements in both rating accuracy and interrater reliability in comparison to a control condition in which neither method was used. Furthermore, even though both methods have the same aim (i.e., enhancing the evaluation process by providing a common evaluative standard for raters), combining both methods led to further improvements in rating accuracy beyond the effects of the individual methods. Practical implications for selection interviews are discussed.

Employment interviews are among the most commonly used procedures for personnel selection worldwide (Buckley, Norris, & Wiese, 2000; Ryan, McFarland, Baron, & Page, 1999). To ensure that the most qualified candidates are selected, it is necessary for interviewees to be

The research reported in this article was supported by a grant from the University of Zurich Research Fund (Forschungskredit der Universität Zürich) to the first author. We want to thank Sabrina Engeli and Patrick Hanselmann for their help with data collection, Leah Hamilton and Stephanie Hastings for their help to improve the writing of the manuscript, and two anonymous reviewers for their constructive suggestions on earlier versions of this paper.

Correspondence and requests for reprints should be addressed to Klaus G. Melchers, Psychologisches Institut der Universität Zürich, Arbeits- und Organisationspsychologie, Binzmühlestrasse 14/12, CH-8050 Zürich, Switzerland; k.melchers@psychologie.uzh.ch.

assessed as accurately and reliably as possible. That said, research has shown that interviewers may differ with regard to how they evaluate information gathered during an interview, and these differences might lower the criterion-related validity of the interview (e.g., Dreher, Ash, & Hancock, 1988; Van Iddekinge, Sager, Burnfield, & Heffner, 2006). To reduce the impact of individual differences between interviewers, it has been suggested that the degree of structure in the interview should be increased. In line with this suggestion, there is widespread agreement that structured interviews have much better criterion-related validity compared to traditional, less structured interviews (e.g., Highhouse, 2008; Huffcutt & Arthur, 1994; Sackett & Lievens, 2008; F. L. Schmidt & Zimmerman, 2004).

As different methods have been advocated to increase the degree of structure in selection interviews (cf. Campion, Palmer, & Campion, 1997; Chapman & Zweig, 2005), the term "structure" has different facets (Macan, 2009). Furthermore, it has been noted that few primary studies have assessed the effectiveness of several proposed methods of increasing interview structure. Specifically, there is a lack of research that demonstrates the unique effects of some of the methods of structure or compares the effectiveness of different methods of structure (Campion et al., 1997). Similarly, more research has been called for that assesses whether or not combining different methods of structure yields additive effects, meaning whether more structure really is better (Dipboye & Gaugler, 1993).

The aim of this study is to provide the first comparison of two methods used to improve interview structure: conducting frame-of-reference (FOR) rater training for interviewers and providing descriptively anchored rating scales for each question. In contrast to descriptively anchored rating scales, evidence for the efficacy of FOR training for interviewers is still missing even though its positive effects have been established for other domains (e.g., for assessment centers [ACs]). Thus, as part of the present comparison, we conduct the first comprehensive evaluation of FOR rater training for interviewers. Furthermore, we develop theoretical arguments that suggest FOR training should be more effective than providing descriptively anchored rating scales even though both methods provide interviewers with a common evaluative standard and thus are conceptually related. Finally, we argue that both methods have disadvantages that might be prevented by combining them, leading us to predict additional improvements in rating quality beyond the effects of the individual methods—a prediction that is in line with the claim that, in interviews, more structure is better. We begin by reviewing relevant empirical and theoretical work and by developing the specific hypotheses tested in this research.

Methods to Improve Interview Structure

The different methods that have been suggested to increase interview structure can be organized according to three dimensions of interview structure recently put forth by Chapman and Zweig (2005). These three dimensions—question consistency, question sophistication, and evaluation standardization—are the result of factor-analytic work that assessed the covariation of different aspects of structure in actual campus interviews and that extended a previous attempt by Campion et al. (1997) to categorize different aspects of structure.

Question consistency captures aspects such as asking the same questions across all candidates, interviewer standardization (i.e., having the same interviewer or interviewer panel conduct all interviews), and basing the interview on a job analysis, so that all questions are linked to actual job requirements. *Question sophistication* includes the use of better interview questions like past-behavior and situational questions, and the limitation of probing and prompting by the interviewer. Finally, *evaluation standardization* consists of such methods as providing descriptively anchored rating scales so that interviewees' answers can be scored against benchmark answers, conducting rater training for interviewers, determining several independent scores, and combining these scores statistically instead of subjectively, meaning according to a predetermined algorithm like unit weighting, for example.

Campion et al. (1997) reviewed theoretical work and past interview research concerning the different methods to improve interview structure as well as research in other fields that is relevant to the interview domain. Campion et al. convincingly argued that the different aspects of structure have positive effects on the quality of the evaluations that interviewers generate during an interview. These positive effects can be seen in improvements in different indicators of rating quality including interrater reliability, rating accuracy (i.e., agreement between interviewers' ratings of an interviewee with externally determined comparison scores in studies in which audio- or videotaped interviews are used), and criterion-related validity. Usually, improvements of one indicator of rating quality are paralleled by improvements of other indicators of rating quality so that improvements of interrater reliability, for example, go hand in hand with improvements of criterion-related validity. Meta-analyses by Conway, Jako, and Goodman (1995) and by Huffcutt and Arthur (1994), for example, used similar classification schemes to code the degree of structure and found improvements in both interrater reliability and criterion-related validity as interview structure increased.

Unfortunately, most of the available meta-analytic research, as well as additional primary studies investigating structured interviews,

confounded different aspects of structure. These confounds make it difficult to assess the relative contributions of the different aspects and to determine whether each of them contributes to the beneficial effects. In addition, even though previous research has confirmed the effectiveness of providing descriptively anchored rating scales, missing primary research on several other important aspects of structure like interviewer training has been noted in the past but this gap has not been filled (cf. Macan, 2009).

Missing knowledge about the effectiveness of different aspects of structure is especially unfortunate when two aspects have the same objective but when one alone might suffice to achieve this objective. This is, for example, the case for providing descriptively anchored rating scales on the one hand and for conducting FOR rater training for interviewers on the other hand. Both methods try to reduce biases in the way that different interviewers evaluate the same information. Given the different paths to reach the same aim, one question is whether both are equally effective in reaching this aim. Theoretical arguments outlined below suggest that this might not be the case. Furthermore, it is unclear whether combining both methods leads to further improvements in rating quality beyond the effects of either method alone. As outlined below, there are conceptual arguments suggesting that a combination of these methods should have beneficial effects. However, given the considerable investment of effort, time, and money involved in implementing these methods to increase interview structure, it is important to gather empirical evidence that supports these claims. Therefore, in this study, we intend to provide answers to these important issues, thereby responding to long-held but still unanswered claims for more research concerning the unique and the combined effects of different methods to increase interview structure (Campion et al., 1997; Dipboye & Gaugler, 1993).

Current Research

In the following sections, we describe relevant theoretical work and review previous research that investigated the aspects of structure that are the focus of the present research—providing descriptively anchored rating scales and conducting rater training for interviewers. Given that the beneficial effects of both methods on the quality of interviewers' evaluations seem rather evident—at least as long as they are implemented appropriately—we try to focus on only the major findings and arguments related to each aspect of structure. However, we will go into more detail when it comes to developing hypotheses concerning the relative and the combined effectiveness of the two methods to improve structure.

Descriptively Anchored Rating Scales

In many structured interviews, descriptively anchored rating scales are used to provide interviewers with scaled examples or descriptions of good, average, and poor answers for each question so that candidates' answers can be evaluated in comparison to preestablished benchmark answers (cf. Campion et al., 1997). Rating anchors reduce rater idiosyncrasies by providing a common evaluative standard to which interviewers can refer each time they are evaluating a candidate's answer. Thereby, descriptively anchored rating scales also limit the memory demands for the interviewers who can refer to the rating anchors as reminders for the different levels of competence and expertise that are required to justify a certain rating. Accordingly, providing such anchored rating scales has long been known to have positive effects on interviewers' interrater reliability (Maas, 1965; Vance, Kuhnert, & Farr, 1978) and accuracy (Vance et al., 1978). Likewise, evidence from one of the few meta-analytic examples where the unique effect of providing anchors was evaluated independent of other aspects of the interviews shows that the use of such anchors increases interviewer reliability and is a crucial factor to ensure good criterion-related validity of selection interviews (Taylor & Small, 2002).

Another impressive and convincing demonstration of the positive effects of providing descriptively anchored rating scales stems from Maurer (2002). He found that ratings made by a sample of naïve raters (undergraduate students) were as reliable and accurate as ratings made by a sample of job-content experts when both samples were provided with the same descriptively anchored ratings scales for the situational interview employed. Furthermore, the student sample provided even more accurate and reliable ratings than another sample of job-content experts who were not provided with such anchors.

Thus, in line with past research, Hypothesis 1 predicts positive effects of using descriptively anchored ratings scales. Specifically, when ratings for interviewees' videotaped answers can be compared with ratings for the same answers from an expert sample, providing anchors should lead to better agreement (i.e., higher accuracy) between participants' ratings and the experts' comparison scores. Thus, we make the following prediction:

Hypothesis 1a: Providing descriptively anchored ratings scales to evaluate interviewees' answers will lead to more accurate ratings in comparison to when no rating anchors are provided.

As noted above, we also expect that descriptively anchored rating scales reduce rater idiosyncrasies by providing a common evaluative

standard for all raters. This should lead to better interrater reliability because different raters should evaluate identical information in a more comparable manner. Furthermore, past research has shown that the effects of descriptively anchored rating scales on interviewer accuracy were paralleled by comparable effects on interrater reliability (e.g., Maurer, 2002; Vance et al., 1978). Therefore, we also predict:

Hypothesis 1b: Providing descriptively anchored ratings scales to evaluate interviewees' answers will lead to better interrater reliability in comparison to when no rating anchors are provided.

Rater Training

Like providing descriptively anchored rating scales, rater training for interviewers aims at enhancing the evaluation process, and several training approaches have been suggested to reach this aim. However, in contrast to the clear evidence concerning the use of improved rating scales, a lack of more systematic research on the effectiveness of interviewer training has long been lamented (Campion et al., 1997; Palmer, Campion, & Green, 1999), and the limited available evidence has produced mixed findings regarding the effects of rater training for interviewers. On the one hand, meta-analytic findings suggest that interviewer training is advantageous and that it contributes to interviewer reliability (Conway et al., 1995) and criterion-related validity (Huffcutt & Woehr, 1999) in real-world selection settings. On the other hand, the promising meta-analytic findings contrast with findings from the few published primary studies on rater training in the interview domain. These primary studies, which were not included in the meta-analyses because they were not conducted in actual selection settings, failed to find evidence for the expected positive effects of rater training on interrater reliability or rating accuracy (Maurer & Fay, 1988; Vance et al., 1978).

There are at least two possible explanations for the diverging results from the meta-analyses versus the primary studies. First, as has been noted in the meta-analyses (cf. Huffcutt & Woehr, 1999), knowledge about the specific content of the interviewer training is often missing so that studies covered by the meta-analyses might have used different training approaches than the primary studies that aimed at evaluating the effects of rater training. Furthermore, because of this missing knowledge, it is also possible that the provision of rater training is confounded with other aspects of structure like asking better interview questions, which is, for example, a topic that is included in many

of the training programs covered in the survey by Chapman and Zweig (2005). Thus, the beneficial meta-analytic effects of providing interviewer training might not (or not only) result from improvements of response standardization.

Second, the failure to find beneficial effects of rater training in the primary studies may well be due to the use of suboptimal training approaches. Specifically, both Maurer and Fay (1988) and Vance et al. (1978) used an approach known as rater error training. Meta-analytic research in the performance appraisal domain has demonstrated that this approach is of limited effectiveness for improving rating accuracy (Woehr & Huffcutt, 1994). Furthermore, some performance appraisal studies even found detrimental effects of this training approach such that trained raters gave less accurate ratings than untrained raters (e.g., Bernardin & Pence, 1980). One aspect of rater error training is that it often places strong weight on teaching raters that their ratings should be normally distributed and that ratings of different dimensions should be independent of each other. Seemingly, this aspect might result in imposing an inappropriate response set on raters who then focus too strongly on these distributional characteristics at the cost of taking ratees' actual performance into account (Bernardin & Pence, 1980).

As an alternative to rater error training, more promising approaches to rater training have been developed in the performance appraisal domain, and these alternative approaches might also be effective for interviewers. Specifically, FOR training seems much more promising as a method to ensure that interviewees' answers are evaluated accurately. Similar to the provision of descriptively anchored rating scales, the main aim of FOR training is to reduce rater idiosyncrasies by enhancing a common understanding of what is required by a given question and by introducing a shared performance theory in general. In contrast to the provision of descriptively anchored rating scales, this is achieved by defining the performance dimensions, defining and describing behavioral examples of different performance levels for each dimension, practicing actual evaluations, and providing feedback to raters concerning the appropriateness of their evaluations prior to the interviews (Bernardin, Buckley, Tyler, & Wiese, 2000).

With regard to the effectiveness of FOR training, we are not aware of any published research investigating its effects in the interview domain. However, previous meta-analytic research in the performance appraisal domain found that FOR training is the most useful training approach to enhance rating accuracy (Woehr & Huffcutt, 1994), and primary studies that explicitly compared FOR training to rater error training found that FOR training was always more effective at improving rating

accuracy than was rater error training (e.g., Bernardin & Pence, 1980; Pulakos, 1984). Furthermore, in the AC domain, FOR training has been found to improve assessor reliability as well as construct- and criterion-related validity of ACs (Lievens, 2001; Schleicher, Day, Mayes, & Riggio, 2002).

In comparison to the performance appraisal or the AC domain, it seems reasonable to expect even stronger effects of FOR training in an interview context because of several important differences that exist between the former domains and the latter. First, raters in the other domains have to deal with actual behavioral observations that require them to correctly interpret potentially ambiguous patterns of behaviors and to relate these patterns of behavior to the targeted rating dimensions. In contrast, when interviewers judge the quality of an answer, they only have to deal with descriptions of behavior that require fewer inferences on their side so that their task is less complex. Second, interview questions often target only a single dimension (e.g., Huffcutt, Weekley, Wiesner, DeGroot, & Jones, 2001; Latham, Saari, Pursell, & Campion, 1980; Van Iddekinge, Raymark, Roth, & Payne, 2006) so that raters do not have to distinguish between different dimensions that a ratee's behavior might reflect, a problem that makes the raters' task more difficult in the other domains. Finally, the observation period in an interview is much shorter and, therefore, the informational basis more constrained and manageable than in ACs or performance appraisals. As such, compared to the cognitive demands placed on individuals evaluating work performance or performance in ACs, the demands placed on interviewers are limited so that it might be easier to overcome potential problems of the rating task by rater training. Taken together, this suggests that, once an interviewer has adopted the appropriate evaluative standard, it is less likely that his or her task will be impeded by the cognitive demands of the rating task.

Nevertheless, previous research on FOR training is limited in that researchers often found stronger effects on some aspects of rating accuracy than on others. Specifically, when ratees are evaluated on different performance dimensions, four different aspects of rating accuracy can be distinguished: Elevation (E), differential elevation (DE), stereotype accuracy (SA), and differential accuracy (DA). These four measures reflect different kinds of discrepancies between participants' ratings and the average ratings of an expert sample (for more information see Murphy & Cleveland, 1995, or Sulsky & Balzer, 1988). E refers to the accuracy of the average rating across all ratees and dimensions. DE refers to accuracy in distinguishing among ratees, averaging across dimensions. SA refers to accuracy with regard to evaluating the different dimensions, averaging across ratees. Finally, DA refers to accuracy in detecting

differences in ratees' specific patterns of strengths and weaknesses across dimensions.

FOR training often produced the strongest effects on aspects of rating accuracy that are related to differentiation between performance dimensions (i.e., on SA and DA) but led to only small or sometimes even no effects on aspects of rating accuracy that are related to the general rating level across ratees and dimensions (i.e., on E; e.g., Day & Sulsky, 1995; Noonan & Sulsky, 2001; Woehr, 1994). This limitation is not so critical for developmental purposes, for example, where information on the strengths and weaknesses of the different ratees is needed to derive appropriate developmental recommendations. However, for selection interviews, where candidates' overall evaluations are relevant to the decision of who should receive a job offer, it is also important that different interviewers hold comparable evaluative standards. This means that E and DE are especially relevant in the present context so that it is important to determine whether FOR training also improves these aspects of rating accuracy.

Given that FOR training aims at establishing a common evaluative standard, training should not only reduce idiosyncrasies concerning evaluations of ratees' specific strengths and weaknesses but also concerning their overall performance. Similarly, training should not only reduce differences between raters but also within raters so that they should also use a more comparable standard across all ratees after the training. Thus, FOR training for interviewers should positively affect all aspects of rating accuracy. Furthermore, given the differences between the various rating domains discussed above suggesting that the rating task in an interview is less demanding than in other rating domains, the chances of finding these effects should be better than in the other domains.

Taken together, we therefore predict that FOR training will lead to improvements of rating accuracy and interrater reliability in interviews:

Hypothesis 2a: FOR training will lead to more accurate ratings compared to when no such training is provided.

Hypothesis 2b: FOR training will lead to better interrater reliability compared to when no such training is provided.

Comparisons of Rater Training and Descriptively Anchored Rating Scales

There are a few arguments that suggest that, compared with conducting FOR rater training, the provision of descriptively anchored rating scales might yield more beneficial effects on the quality of interviewers' ratings. First, as noted above, interviewers can refer to descriptively anchored rating scales each time they have to evaluate an answer during

the interview. Thus, the demands on interviewers' memory are reduced in contrast to rater training where impairments of rating quality can occur once an interviewer forgets about relevant aspects of the performance theory. And second, meta-analytic evidence by Huffcutt and Woehr (1999) revealed that the level of structure, which captured providing descriptively anchored rating scales, was more closely related to criterion-related validity than the provision of interviewer training. However, Huffcutt and Woehr's coding for the level of structure in this meta-analysis encompassed providing rating anchors as well as the use of a predetermined set of questions for a given primary study. Thus, it was not possible to evaluate separately the effects of providing anchors. Furthermore, as far as we are aware, no primary research has compared the effects of the two methods of structure.

Despite the above-mentioned arguments suggesting that descriptively anchored rating scales might be more effective than the provision of rater training, there are at least three reasons that lead us to predict that rater training may be the superior of the two methods. First, evidence from the performance appraisal domain suggests that practicing actual evaluations and receiving feedback about the appropriateness of one's evaluations is an important component of FOR training that is missing when interviewers are only provided with descriptively anchored rating scales. Research by Sulsky and Kline (2007) and by Athey and McIntyre (1987), for example, found that feedback concerning the accuracy of the trial evaluations contributes to the positive effects of FOR training.

Second, it was found that FOR training leads to greater agreement with the performance theory taught during training (Schleicher & Day, 1998) and that agreement with this performance theory is related to more accurate ratings (Schleicher & Day, 1998; Uggerslev & Sulsky, 2008). Discussions elicited during FOR training in exercises that ask participants, for example, to assign behavioral examples for the different performance dimensions to different levels of performance probably foster the adoption of the performance theory. Similarly, the discussions during the provision of feedback for the trial evaluations probably have a similar effect because in both cases they help to clarify difficulties of the performance theory and potentially to overcome resistance on the side of the training participants. Again, these beneficial effects are absent when only descriptively anchored rating scales are provided.

And third, at a theoretical level, it has been argued that rating practice and feedback contribute to an elaborative rehearsal of the training information. According to levels-of-processing theory (Craik, 2002; Craik & Lockhart, 1972), deeper levels of processing that involve more elaborative cognitive analyses will result in a richer and more detailed memory

trace and better retention of the information learned. The exercises and the discussion and feedback provided during FOR training are likely to contribute to more elaborative processing. Again, however, these aspects are missing when only descriptively anchored rating scales are provided.

Given that practice and feedback are important for the effectiveness of FOR training, some additional discussion seems warranted in light of research from other domains that revealed that too frequent or too specific feedback can be disadvantageous for transfer of learned knowledge or skills. In the domains of motor learning or cognitive learning, for example, R. A. Schmidt and Bjork and their respective colleagues found that too frequent feedback can impair long-term retention or transfer even though this feedback has beneficial effects on performance during the training phase (cf. the research reviewed by R. A. Schmidt & Bjork, 1992). Similarly, research by Goodman and Wood (2004, 2009) and Goodman, Wood, and Hendrickx (2004) in organizational behavior also revealed that highly specific feedback can have positive effects on training performance but no or even negative effects on performance in a transfer test.

Several explanations have been put forth for these disadvantageous effects of feedback, such as the learner becoming too dependent on this informational support, feedback attracting too much attention and thereby interfering with the learner's own cognitive processing, or feedback negatively affecting exploration during learning. However, it seems unlikely that feedback during FOR training has these disadvantageous effects because several aspects of FOR training try to prevent such negative effects: First, the design and selection of the training materials usually ensure that trainees experience examples from the entire range of performance; second, feedback is usually provided in a manner that tries to prevent cognitive overload on the side of the trainees; and third, feedback is usually given to the entire group of trainees so that individual trainees do learn not only from their own trial evaluations but also from the trial evaluations made by their fellow trainees. Furthermore, in addition to providing information on the correctness of these trial evaluations, feedback during FOR training also has the function to open up discussions between trainees about different raters' reasons for assigning their respective ratings. Besides fostering more elaborative processing, these discussions are a necessary means to ensure that different raters hold comparable evaluative standards at the end of the training. Finally, this discussion also provides opportunities to provide additional instruction and to clarify potential misunderstandings or uncertainties.

Taken together, because of the beneficial aspects of FOR training outlined above that are missing in the case of providing descriptively anchored rating scales, we expect that FOR training is more effective at reducing rater idiosyncrasies. This, in turn, should lead to more accurate

ratings. Furthermore, better adoption of a common evaluative standard across raters should also lead to higher interrater reliability when FOR training is conducted than when descriptively anchored rating scales are provided. Therefore, we make the following predictions:

Hypothesis 3a: FOR rater training will lead to more accurate ratings than the provision of descriptively anchored rating scales.

Hypothesis 3b: FOR rater training will lead to better interrater reliability than the provision of descriptively anchored rating scales.

Combined Effects of Rater Training and Descriptively Anchored Rating Scales

Even though there is general agreement that structure is good for interview reliability and validity, an important question is whether more structure is better. In fact, some researchers have expressed skepticism regarding the effectiveness of combining different aspects of structure (Campion et al., 1997). Furthermore, in the present context, a combination of providing descriptively anchored rating scales and FOR training might yield no additional gains in rating accuracy and interrater reliability when a common evaluative standard is already established for raters.

However, meta-analytic evidence suggests that interviewer training can explain incremental validity beyond the effects of other aspects of structure including the provision of descriptively anchored rating scales (Huffcutt & Woehr, 1999). In line with this, Huffcutt and Woehr recommended that training should be provided to interviewers regardless of whether or not the interview itself (i.e., the questions and rating scales) is structured. Yet, as noted above, it should be recalled that the provision of training in this meta-analysis was potentially confounded with aspects such as asking better interview questions. However, research in the performance appraisal and the AC domains also found that FOR training improved rating accuracy even when descriptively anchored rating scales were used (e.g., Schleicher et al., 2002; Sulsky & Day, 1994).

A potential reason for the promising empirical findings is that employing both methods to increase structure combines their respective advantages and prevents their respective disadvantages. On the one hand, as explained in the previous section, FOR training provides opportunities for practice, feedback, and discussion, which together contribute to better adoption of the relevant performance theory and to deeper processing of relevant information. On the other hand, providing descriptively anchored rating scales might reduce the need for recall of all relevant aspects of

the performance theory because the anchors provide memory cues that are always present when raters make their evaluations. Previous findings that memory for training content correlated with rating accuracy (Noonan & Sulsky, 2001; Roch & O'Sullivan, 2003; Sulsky & Kline, 2007) suggest that descriptively anchored rating scales might help to reduce impairments of rating quality that might occur when raters forget important information. And even though we argued that FOR training is more effective in the interview domain than in other domains, it might not suffice to completely reduce rater idiosyncrasies so that the additional provision of rating anchors might help to overcome the disadvantages of rater training. Accordingly, we make the following predictions:

Hypothesis 4a: Combining FOR rater training and providing descriptively anchored rating scales will lead to more accurate ratings than either method alone.

Hypothesis 4b: Combining FOR rater training and providing descriptively anchored rating scales will lead to better interrater reliability than either method alone.

Summary of Goals of the Present Study

This study presents a comparison of the effects of two important methods to increase structure in selection interviews—conducting FOR rater training and providing descriptively anchored rating scales—on rating accuracy and interrater reliability. By doing so, we also present a comprehensive evaluation of FOR training in the interview domain. Finally, this study evaluates the hypotheses that a combination of both methods increases rating accuracy and interrater reliability beyond the effects of either method alone because using both methods should combine their respective advantages and prevent their respective disadvantages. Thereby, our research also provides guidance concerning the question of whether it is worth going the extra mile and combining these two methods in actual selection settings, as they both require considerable investments of time and money.

Method

Participants

We used a 2 × 2 between-subjects design in which we assessed two factors: rater training (FOR training vs. control training) and rating anchors (descriptively anchored scales vs. no descriptively anchored scales). Altogether, 199 participants took part in this study. Of these

participants, 49 were men and 150 were women, and their age varied between 18 and 47 years, with a median of 23. The vast majority of the participants were psychology undergraduates who were either recruited from several big lectures from different areas of psychology or registered in the department's database of study participants. Most were paid for their participation but some also participated to meet a course requirement.

Development of Stimulus Materials

For the study we used standardized stimulus materials. The development of these materials proceeded in several steps. First, seven situational interview questions were taken from a longer interview developed for a study by Klehe, König, Richter, Kleinmann, and Melchers (2008, Study 2) and were suitable for university graduates who are applying for a management trainee position. The procedure used to develop the original interview is described in detail by Klehe et al. and in brief consisted of the following steps: identification of potentially job-relevant dimensions on the basis of critical incidents, determining which of the dimensions are suitable to be assessed in a structured interview, selection of a subset of dimensions that should be most independent from each other, and collecting and selecting potential questions for the different dimensions. Most of these questions were taken from structured interviews developed and validated for the banking (e.g., Schuler & Moser, 1995) and engineering sectors (Deller & Kleinmann, 1993). Klehe et al. then pretested whether the questions indeed reflected the targeted dimensions and whether the descriptive rating anchors provided for each question reflected poor, average, or good answers. The final set of questions chosen for this study was intended to tap into systematic planning (defined as correctly prioritizing tasks, structuring of tasks and projects, planning ahead of time, allocating tasks, and setting goals), cooperation (defined as considering others' needs and assisting with problems that they may have as well as being prepared to compromise with others), and leadership (defined as striving for and taking on responsibility for tasks and groups, coordination of teams, and arguing one's point of view within a group). Two of the questions targeted systematic planning, two targeted leadership, and three targeted cooperation. A sample question together with its corresponding descriptively anchored rating scale is provided in the Appendix.

We developed seven different scripts with answers to the entire set of interview questions and then filmed these interviews. Each script consisted of one answer to each of the seven interview questions. As the basis for these scripts, we used notes that interviewers had taken during actual interviews in studies in which the interview questions were used (Klehe et al., 2008; Melchers et al., 2009). For each script, we included answers that reflected different levels of performance according to the

interviewers' ratings as well as our own judgment of the content of the answers. Furthermore, across the different scripts we also ensured that the answers to a given question varied in content and performance level. Finally, we used a different interviewee for each of the videos. The length of each videotaped interview was approximately 7 minutes.

Next, we followed procedures described by Sulsky and Balzer (1988) to determine comparison scores for the interviewees' answers. Seven experts judged each of the interviewees' answers. Three of the experts held PhDs in the field of work and organizational psychology, and the other four were master's level work and organizational psychology students. All were familiar with the interview questions, had undergone interviewer training in the past, and had practical experience in conducting interviews in which the current seven situational questions were used. The experts were given the written interview questions as well as the descriptively anchored rating scales that were used in the study by Klehe et al. (2008). They watched the videotapes on their own and could stop and rewind as often as they wished. After each video, the experts also rated the perceived realism of the interview on a seven-point scale ranging from 1 = *completely unrealistic* to 7 = *very realistic*.

To determine the final comparison scores for each videotaped interview, we averaged the experts' ratings for each answer. Before doing so, we determined the experts' interrater agreement for each question in each of the videotaped interviews by calculating two common indices of agreement (LeBreton & Senter, 2008), r_{wg} (James, Demaree, & Wolf, 1984) and the AD index (Burke, Finkelstein, & Dusig, 1999). r_{wg} assesses agreement by comparing the variance obtained from multiple raters to the variance one might obtain if the ratings were due to random responding. In the present case, random responding was assumed to be reflected by a uniform response distribution. Values for r_{wg} can vary between 0 and 1 with larger values indicating better agreement. In contrast, values for AD estimate agreement in the metric of the original scale by determining the absolute deviation of each rating from the mean of the group rating (i.e., the mean expert rating) and then averaging the deviations. Accordingly, smaller values of AD indicate better agreement.

For this study, we selected one video to be used for the training sessions (Video T) and three target videos for which participants later had to rate the interviewees' answers after their training (Videos 1 to 3). We selected these four videos on the basis of the experts' interrater agreement and the average realism ratings for each video. The mean interrater agreement values across the different questions and the realism ratings for the selected videos are shown in Table 1. This table also contains the mean expert scores for the four videos. All mean interrater agreement values are better than the critical values for statistical significance as well as the cut-off scores for practically useful levels of interrater agreement

TABLE 1
Interrater Agreement, Realism Ratings and Expert Scores for Interview Videos

	Video T	Video 1	Video 2	Video 3
r_{wg}	0.72	0.93	0.81	0.82
AD	0.50	0.23	0.33	0.37
Realism	6.00	5.86	5.57	5.57
Question 1	4.71	4.86	3.86	2.71
Question 2	3.00	5.00	3.86	1.14
Question 3	3.57	1.29	1.00	3.00
Question 4	1.71	1.57	3.86	1.00
Question 5	5.00	5.00	3.86	1.57
Question 6	1.42	3.14	5.00	1.00
Question 7	2.86	4.14	1.14	3.14

Note. r_{wg} is James et al.'s (1984) within-group interrater agreement index, AD is Burke et al.'s (1999) average deviation index (in both cases, the values reflect the mean agreement across all seven questions). Realism ratings and expert scores are based on averaged data from all $n = 7$ expert raters. Video T was later used for the training and Videos 1 to 3 were later used for the test.

provided by Dunlap, Burke, and Smith-Crowe (2003; also see Burke & Dunlap, 2002): For five-point measurement scales and seven raters as in this case, values for r_{wg} that are larger than .67 and .70, respectively; are statistically significant; and practically useful. Similarly, values for the AD index that are smaller then .61 and .83, respectively, are statistically significant and practically useful. Thus, the levels of interrater agreement obtained justify the use of the mean expert ratings as comparison scores. Finally, the average realism ratings for the videos ranged from 5.57 to 6.00 on the seven-point scale, indicating that the filmed interviews were regarded as realistic.

Procedure

Groups of participants were randomly assigned to one of the four experimental conditions. Between 4 and 12 participants took part in each training session (see below), each of which was conducted by two trainers together. Each training session lasted for about 70 minutes. After the training session had finished, participants had to fill in a questionnaire unrelated to the present study. This questionnaire was intended as a brief distractor task before the final test stage, and participants needed approximately 10 minutes to complete it.

Interview Rating Forms and Descriptively Anchored Rating Scales

For each interview question, a five-point rating scale ranging from 1 = *poor* to 3 = *average* to 5 = *good* was used to evaluate the interviewees'

answers. The rating scales were printed below each interview question and were labeled by the dimension that the respective question was intended to assess. Space was provided between each interview question and the respective rating scale so that participants could take notes.

Half of the participants were provided with descriptively anchored rating scales. These benchmark anchors described what actions an interviewee might describe for a poor, an average, and a good answer. The descriptively anchored rating scales, which were taken from the study by Klehe et al. (2008), were the same as those provided to the expert sample described above. The remaining participants were only provided with a graphic rating scale containing the labels *poor*, *average*, and *good* at the scale values 1, 3, and 5, respectively. The graphic rating scales did not contain descriptive anchors.

Rater Training

Half of the participants who were later provided with descriptive rating anchors for the rating task took part in the FOR training and the other half took part in the control training. Similarly, half of the participants who were later only provided with the graphic rating scales took part in the FOR training and the other half took part in the control training, so that both factors were administered in a fully crossed design.

Frame-of-reference training. The aim of the FOR training was to impose a common evaluative standard on participants. In a brief lecture, participants were first introduced to the difference between structured versus unstructured interviews and received information about the target position for which the interview was developed. They were then introduced to the different target dimensions. They were given a sheet of paper on which the dimensions were defined and explained by behavioral examples. The trainers read through the definitions and behavioral examples, explained them, and answered participants' questions.

After having read through the dimension definitions, participants were presented with an exercise in which they received the interview questions one by one accompanied by three written statements that corresponded to poor, average, and good answers, respectively. These statements were paraphrases of the descriptive anchors for the separate questions. Participants had to assign each statement to the correct performance level. After each question, the trainers discussed the participants' suggestions with the group, provided feedback concerning the accuracy of these assignments, and explained why the written answers corresponded to the respective performance level.

After participants had completed the exercise, they were handed the interview forms on which all interview questions and the ratings scales were printed. As noted above, for half of the FOR training groups, these

rating scales contained descriptive anchors, and for the other half they did not. Finally, participants were shown Video T and had to score the answers to all seven questions. Afterwards, the trainers elicited a discussion of how the participants had decided upon an assigned rating, gave them feedback with regard to how the videotaped interviewee was evaluated by the experts, and clarified any discrepancies between the participants' and the experts' ratings.

Control training. We employed a pseudotraining condition instead of a no-training control condition. In line with previous rater training studies, this pseudotraining condition familiarized participants with the stimulus materials and provided participants with information that was related to their rating task but did not contain the critical elements of FOR training (e.g., Roch & O'Sullivan, 2003; Schleicher et al., 2002; Sulsky & Day, 1994). The rationale for this was to ensure that participants in the control training conditions were equally familiar with the training context, the stimulus materials, and other structural aspects of the training situation (cf. Stamoulis & Hauenstein, 1993). Thereby this condition prevents control participants from providing less accurate ratings just because of their lower familiarity with the task and therefore provides a more conservative test of the potential effects of rater training. Thus, the control training was of comparable length and was kept as similar as possible to the FOR training, with the following exceptions: The lecture also included information about different selection techniques and about possible ways to improve selection interviews by combining different types of questions and other interview components in what is known as a multimodal interview (cf. Schuler & Funke, 1989). As part of the information about different types of interview questions, participants were introduced to situational questions and were informed that the later interviews they had to evaluate would consist of such questions. Furthermore, the lecture included information about the critical incident technique as a way to gather job information and to develop interview questions. Finally, this part of the training contained an exercise in which participants had to formulate a situational interview question on the basis of a critical incident provided by the trainers.

Next, participants were asked to read the written interview questions and encouraged to ask for clarification if they did not understand any of the interview questions. In contrast to the FOR training, participants did not receive definitions or behavioral examples of the dimensions. Similarly, they were not provided with the exercise in which they had to assign the written answers to the appropriate rating categories.

Finally, participants were handed the interview forms on which all interview questions and the ratings scales were printed. Again, for half of the control training groups, these rating scales contained descriptive

anchors, and for the other half they did not. Then participants were shown the same videotaped interview as in the FOR training and had to score each answer. In contrast to the FOR training, however, the trainers neither discussed how participants had decided upon an assigned rating nor did they provide feedback about the accuracy of the participants' ratings.

Test Stage

In the test stage, participants were shown Videos 1 to 3 and had to evaluate each interviewee on each answer. Across participants, we counterbalanced the order in which the videos were shown. Finally, the participants answered a brief questionnaire containing demographic questions as well as some items that were used as control variables.

Variables

Control variables. The final questionnaire included several items used to assess the participants' motivation to provide accurate ratings and the difficulty of evaluating the interviewees' answers. These items were answered on Likert scales ranging from 1 to 5 with higher numbers indicating stronger agreement. We included accuracy motivation to enable an assessment of whether participants who took part in the control training were less motivated to provide accurate evaluations, for example, because they might have noticed that the training did not provide them with the necessary evaluative standards they needed to evaluate the candidates' answers. This was important because previous research has found that rater motivation can influence the accuracy of their ratings (e.g., Salvemini, Reilly, & Smither, 1993). Accuracy motivation was measured with two items specifically developed for this study: "I was motivated to evaluate the candidates as appropriately and accurately as possible" and "It was important to me to evaluate the candidates in a fair manner."

To assess whether training as well as the provision of descriptively anchored rating scales actually facilitated the rating task in the eyes of the participants, we included two items to assess the difficulty of evaluating the interviewees' answers. These items ("I found it difficult to evaluate the candidates as appropriately and as accurately as possible" and "It was easy to evaluate the quality of the different answers") were also developed specifically for this study.

Rating accuracy. As dependent variables, we employed the four indicators of rating accuracy already mentioned in the Introduction: E, DE, SA, and DA. E refers to the accuracy of the average

rating across all $n = 3$ interviewees and all $k = 7$ questions.[1] DE refers to accuracy in distinguishing among interviewees, averaging across questions. SA refers to accuracy with regard to evaluating the different interview questions, averaging across interviewees. Finally, DA refers to accuracy in detecting differences in interviewees' specific patterns of strengths and weaknesses across questions.

These accuracy scores may be expressed by the following equations:

$$E^2 = (\overline{x}.. - \overline{t}..)^2 \tag{1}$$

$$DE^2 = 1/n \sum_i [(\overline{x}_i. - \overline{x}..) - (\overline{t}_i. - \overline{t}..)]^2 \tag{2}$$

$$SA^2 = 1/k \sum_j [(\overline{x}._j - \overline{x}..) - (\overline{t}._j - \overline{t}..)]^2 \tag{3}$$

$$DA^2 = 1/kn \sum_i \sum_j [(x_{ij} - \overline{x}_i. - \overline{x}._j + \overline{x}..) - (t_{ij} - \overline{t}_i. - \overline{t}._j + \overline{t}..)]^2 \tag{4}$$

where x_{ij} and t_{ij} = rating and comparison score for interviewee i on question j; $\overline{x}_i.$ and $\overline{t}_i.$ = mean rating and mean comparison score for interviewee i across all questions; $\overline{x}._j$ and $\overline{t}._j$ = mean rating and mean comparison score for question j across all interviewees; and $\overline{x}..$ and $\overline{t}..$ = mean rating and mean comparison score across all interviewees and questions. As all accuracy measures are calculated as squared deviations between the participants' evaluations and the comparison scores from the expert raters (Sulsky & Balzer, 1988), smaller values indicate better accuracy.

As noted above, for selection interviews where candidates' overall evaluations are relevant to decide who should receive a job offer, it is important that interviewers use a comparable evaluative standard across different candidates and that different interviewers also hold comparable

[1] In performance appraisal research the different accuracy measures are usually calculated on the basis of ratings of different candidates on several different dimensions. In our study, we had multiple questions for each targeted interview dimension so that we used the single questions as the basis to determine rating accuracy. However, additional analyses, for which we first averaged ratings for all questions that targeted the same dimension and then used these dimension scores, always led to comparable results and conclusions with regard to our hypotheses and research questions.

TABLE 2
Means, Standard Deviations, and Intercorrelations for Study Variables for the Whole Sample (N = 199)

Variable	*M* (*SD*)	1	2	3	4	5	6
1. Motivation	4.78 (0.40)						
2. Difficulty	2.49 (0.83)	.04					
3. E	0.08 (0.12)	–.03	.09				
4. DE	0.09 (0.12)	–.16*	.16*	.24**			
5. SA	0.27 (0.17)	–.03	.30**	.40**	.30**		
6. DA	0.45 (0.31)	–.10	.20**	.30**	.55**	.45**	

Note. Smaller values for accuracy variables indicate better accuracy.
*$p < .05$, **$p < .01$.

evaluative standards. Thus, in the present context, E and DE are especially relevant.

Interrater reliability. Finally, for each interview, we calculated an intraclass correlation (ICC 2.1; Shrout & Fleiss, 1979) as an indicator of the reliability of the participants' ratings and then averaged the three intraclass correlations across interviews. This intraclass correlation takes into account the consistency between raters, meaning whether different raters produce the same rank-order of the rated answers, as well as the consensus between raters, meaning whether different raters agree with regard to the absolute ratings of the rated answers (LeBreton & Senter, 2008; McGraw & Wong, 1996).

Analytical Strategy

All hypotheses concerning rating accuracy were evaluated by using A one-way analyses of variance (ANOVAs) followed by planned contrasts that addressed the specific comparisons of interest. Given the number of comparisons, we considered Bonferroni-adjusted levels of significance to prevent an inflation of Type 1 errors. Thus, $p = .0083$ was used as the level of significance for all contrasts. Furthermore, all hypotheses concerning interrater reliability were evaluated by comparing the respective interrater reliabilities.

Results

Means, standard deviations, and correlations between all study variables for the whole sample can be found in Table 2. Furthermore, Table 3 shows means and standard deviations separately for the different experimental groups and also contains interrater reliabilities.

TABLE 3
Means and Standard Deviations for the Different Experimental Groups

	Control training		FOR training	
	No anchors $n = 51$ *M* (*SD*)	Anchors $n = 50$ *M* (*SD*)	No anchors $n = 52$ *M* (*SD*)	Anchors $n = 46$ *M* (*SD*)
Motivation	$4.78\ (0.38)_{a}$	$4.74\ (0.39)_{a}$	$4.77\ (0.40)_{a}$	$4.82\ (0.45)_{a}$
Difficulty	$2.98\ (0.93)_{a}$	$2.18\ (0.75)_{b}$	$2.55\ (0.77)_{c}$	$2.23\ (0.59)_{b}$
E	$0.17\ (0.18)_{a}$	$0.06\ (0.07)_{b}$	$0.06\ (0.08)_{b,c}$	$0.03\ (0.04)_{c}$
DE	$0.17\ (0.18)_{a}$	$0.08\ (0.10)_{b}$	$0.05\ (0.06)_{b}$	$0.05\ (0.05)_{b}$
SA	$0.45\ (0.20)_{a}$	$0.20\ (0.10)_{b,c}$	$0.23\ (0.13)_{b}$	$0.17\ (0.08)_{c}$
DA	$0.75\ (0.34)_{a}$	$0.37\ (0.27)_{b}$	$0.41\ (0.20)_{b}$	$0.26\ (0.13)_{c}$
Interrater reliability	$.40_{a}$	$.71_{b}$	$.72_{b}$	$.78_{b}$

Note. Smaller values for accuracy variables indicate better accuracy. Means of the control variables and the accuracy measures in the same row that do not share a common subscript differ according to planned contrasts using Bonferroni-adjusted levels of significance.

Preliminary Analyses

The means for participants' motivation in all four conditions were close to the maximum score of 5, meaning that all participants were well motivated to provide accurate ratings (cf. Table 3). This was confirmed by a one-factorial ANOVA comparing the four groups that did not reveal any differences between them, $F < 1$. Differences in the accuracy variables can therefore not be attributed to any differences in participants' motivation to provide accurate ratings.

In contrast, the groups differed with regard to how difficult they found the rating task so that the corresponding ANOVA found a significant main effect between the four groups, $F(3, 195) = 11.30$, $p < .01$, $\eta^2 = .15$. Subsequent planned contrasts revealed that the group that only participated in the control training and that was not provided with rating anchors found the task more difficult than the groups that were not provided with rating anchors (subscripts in Table 3 indicate significant differences). Thus, the provision of rating anchors made the task easier for participants.

Test of Hypotheses

As noted above, means, standard deviations, and interrater reliabilities can be found in Table 3. It should be noted that larger values for the accuracy scores indicate lower accuracy, as the accuracy scores are defined

as deviations from the experts' mean scores. Again, subscripts for the means in a row indicate significant differences.

One-factorial ANOVAs revealed that the four experimental groups differed significantly with regard to each of the accuracy scores: $F(3, 195) = 15.48$, $\eta^2 = 19$, for E; $F(3, 195) = 12.40$, $\eta^2 = 16$, for DE; $F(3, 195) = 42.38$, $\eta^2 = .39$, for SA; and $F(3, 195) = 34.83$, $\eta^2 = 35$, for DA, all ps $< .01$. Results for the specific hypotheses are presented in detail in the following paragraphs.

Hypotheses 1a and 1b predicted that the provision of descriptively anchored rating scales would lead to better rating accuracy and higher interrater reliability. In line with these predictions, of the two groups that had only taken part in the control training, the group that was provided with descriptive anchors had lower accuracy scores (i.e., showed better accuracy) than the comparison group for all four accuracy measures, and all of the Bonferroni-adjusted contrasts turned out to be significant. The effect sizes for the comparisons (expressed as Cohen's *d*s) were 0.75, 0.63, 1.54, and 1.21 for E, DE, SA, and DA, respectively. Thus, most of the differences reflect large effects (Cohen, 1992). Finally, the results concerning interrater reliability paralleled those for the accuracy scores meaning that the group that was provided with descriptive anchors had markedly better reliability (.71) than the group without descriptive rating anchors (.40). Again, this difference turned out to be significant, $z = 2.26$, $p < .05$. Taken together, our results provide good support for Hypotheses 1a and 1b.

Hypotheses 2a and 2b predicted that FOR rater training would lead to better rating accuracy and higher interrater reliability. In line with these hypotheses, of the groups that were not provided with descriptively anchored ratings scales, the group that had taken part in FOR training had better accuracy (i.e., lower scores) than the group that had taken part in the control training; the contrasts for all four accuracy measures turned out to be significant (cf. Table 3). The effect sizes for the comparisons were 0.76, 0.85, 1.30, and 1.21 for E, DE, SA, and DA, respectively. Thus, all differences reflect large effects. Similarly, concerning interrater reliability, the FOR group had a markedly higher value (.72) than the group without FOR training and descriptive rating anchors (.40). Again, the difference turned out to be significant, $z = 2.38$, $p < .05$. Thus, our results support Hypotheses 2a and 2b.

Hypotheses 3a and 3b predicted that FOR training leads to more accurate and reliable ratings than the provision of descriptive anchors. However, our results did not reveal significant differences for any of the accuracy scores between the group that was provided with FOR training but not with descriptive anchors and the group with control training and

descriptive anchors. Accordingly, the absolute values for the effect sizes ranged only between .00 and .27. The results for the accuracy scores paralleled the nearly identical reliabilities obtained (.72 and .71). Thus, the present results do not support Hypotheses 3a and 3b.

The final set of comparisons concerned Hypotheses 4a and 4b, predicting that a combination of both aspects of structure leads to more accurate and reliable ratings than either aspect alone. In line with these predictions, examination of the means in Table 3 shows that the group that could benefit from FOR training as well as from the provision of descriptively anchored rating scales always had the lowest (i.e., the most accurate) accuracy scores. Concerning the comparison with the FOR group that was not provided with descriptive anchors, significant results were found for two of the four planned contrasts for the accuracy scores. These differences reflected medium to large effect sizes of 0.53 and 0.88 for SA and DA. Furthermore, the contrast for E also reflected a medium-size difference of 0.51 but fell just short of significance ($p = .01$) after the Bonferroni adjustment, which would have required a significance level of $p = .0083$ (for DE, both groups had the same mean). When we compared the group that could benefit from both methods to improve structure with the group that was only provided with descriptive anchors, we found differences in the small to medium range of 0.58, 0.32, 0.37, and 0.55 for E, DE, SA, and DA, respectively. The planned contrasts revealed significant differences for E and DA.

In addition to the planned contrasts, we also conducted 2×2 ANOVAs (Training $\times$ Anchors) for each accuracy score to evaluate the interplay of the two aspects of structure. For each accuracy score, these ANOVAs revealed significant main effects for conducting FOR rater training and for providing descriptive anchors, all Fs$(1, 195) > 9.21$, all ps $< .01$. The η^2s for FOR training ranged between .08 and .17, and the η^2s for providing anchors ranged between .04 and .19. Furthermore, the interaction terms were significant for all accuracy scores, all Fs$(1, 195) > 7.96$, all ps $<$ .05, η^2s between .02 and .07. Together with the results from the planned contrasts, the significant interactions indicate that combining both aspects of structure leads to more accurate ratings than providing only one aspect of structure but that these beneficial effects are not additive but smaller than the effects found in relation to Hypotheses 1a and 2a.

Finally, with regard to interrater reliability, the group that could benefit from both aspects of structure descriptively provided more reliable ratings than the groups that could benefit from only one aspect (.78 vs. 72 and .71, respectively). However, none of the differences was significant. Thus, our results support Hypothesis 4a but not Hypothesis 4b.

Discussion

This study represents the first investigation to compare the effects of providing descriptively anchored rating scales and conducting FOR rater training with regard to their effectiveness in improving the accuracy and interrater reliability of ratings in a structured interview. In line with Hypotheses 1a and 1b, the use of descriptively anchored rating scales improved rating accuracy and interrater reliability for participants who had only taken part in control training. These findings are consistent with previous interview research (e.g., Maas, 1965; Maurer, 2002; Vance et al., 1978). Similarly, and in line with Hypotheses 2a und 2b, FOR training improved rating accuracy and interrater reliability for participants who were not provided with descriptively anchored rating scales. Thus, in contrast to past research in the interview domain (e.g., Maurer & Fay, 1988; Vance et al., 1978) that only evaluated so-called rater error training and failed to find improvements of rating quality, this study is the first primary study to provide clear evidence that appropriate training (i.e., FOR-training, cf. Woehr & Huffcutt, 1994) can make a marked difference with regard to the quality of interview evaluations. In this way, this study also responds to long-raised but still unanswered calls for more research concerning the actual effects of interviewer training (Campion et al., 1997; Palmer et al., 1999).

The results concerning the effectiveness of FOR training are also important because they revealed that training substantially improved aspects of rating accuracy that reflect the general rating level across ratees and dimensions or comparable rating levels across ratees (i.e., E and DE). This is important because past rater training research often found no or only small effects on these aspects (e.g., Day & Sulsky, 1995; Noonan & Sulsky, 2001; Woehr, 1994). As argued above, for selection interviews in which overall evaluations of candidates are relevant to decide who should receive a job offer, it is also important that all interviewers hold comparable evaluative standards and that an interviewer does not use different evaluative standards across different candidates. In addition, even though we found stronger effects on SA and DA, the effect sizes for E and DE were also large according to conventional standards (Cohen, 1992). Furthermore, concerning the size of the training effects, we also consider our results to be supportive of our reasoning from above that FOR training leads to stronger effects in the interview domain than in the performance appraisal or the AC domain. Research in these domains often found smaller effects than this study (e.g., Day & Sulsky, 1995; Schleicher et al., 2002; Woehr, 1994).

Compared with one another, the two methods to improve structure were comparable with regard to rating accuracy and interrater reliability. Thus, the present results suggest that the two methods not only aim at the same objective (i.e., improving rating quality by providing a common evaluative standard for raters) but also achieve this objective to a similar degree. This is in contrast to Hypotheses 3a and 3b predicting that FOR training would be more effective in improving rating accuracy and interrater reliability. Thus, the assumed positive effects of providing feedback to trainees, of fostering greater agreement with the performance theory, or of leading to deeper levels of processing did not lead to more accurate and reliable ratings. An additional interesting finding, however, was that although both components of structure seemed equally effective with regard to improving rating accuracy, providing descriptively anchored rating scales seemed to have a stronger effect on lowering the perceived difficulty of the rating task than having participated in FOR training. Thus, it might be possible that descriptively anchored rating scales can outweigh the beneficial effects of FOR training by indeed lowering the demands on interviewers' memory. Specifically, because the anchors provide memory cues that interviewers can use during the interview, there is a reduced need for recall of all relevant aspects of the performance theory.

With regard to the question of whether the combination of providing behavioral rating anchors and FOR training to evaluate answers in a structured interview increases rating quality beyond the effects of either method alone (Hypotheses 4a and 4b), our results suggest that more structure really is better. Specifically, the group that participated in FOR training and was provided with rating anchors descriptively showed the strongest effects for each of the four accuracy measures and for interrater reliability. And even though the comparisons with the two groups that could only benefit from one aspect of structure did not reveal a significant difference concerning interrater reliability, we found several significant differences for the accuracy scores, which lend support to Hypothesis 4a. The finding that a combination of both aspects of structure can lead to further improvements of rating quality—especially concerning E—is important because it mitigates previous concerns (Campion et al., 1997) that either component alone is sufficient to establish the common evaluative standard that is needed to substantially reduce biases otherwise introduced by rater idiosyncrasies. However, our findings are in line with the idea that using both methods to improve structure combines their respective advantages and prevents their respective disadvantages. This corroborates recommendations by Huffcutt and Woehr (1999) to provide training to interviewers regardless of whether or not the interview itself (i.e., the questions and rating scales) is structured.

Limitations and Suggestions for Future Research

A limitation of this study is that it only looked at short-term effects. This means that participants were only tested directly after training. Even though we used a brief 10 minute distractor task between training and testing, the finding that the two methods to improve structure are of comparable effectiveness might be specific to the time of testing. If participants are tested later after training, it might well be that training effects become blurred whereas the effects of providing descriptively anchored rating scales remain the same, as these scales provide a frame of reference that does not make any demands on memory for training content because they are always available at the time of the rating task. Furthermore, past research has found that some features of the training design are beneficial for performance during or directly after training but are not related to retention or transfer to new materials or situations (cf. R. A. Schmidt & Bjork, 1992).

Given the possibility that the reduced demands on raters' memory are responsible for not finding a difference between FOR training and descriptively anchored rating scales in this study, more information about retention of rating skills acquired during rater training and the long-term effects of this training is needed. Unfortunately, research on the stability of the effects of rater training over time is sparse. Even though the few studies that exist (Roch & O'Sullivan, 2003; Sulsky & Day, 1994) found no significant deterioration of training effects over time, these results should be regarded with caution because of the short delays considered (Sulsky & Day, 1994, for example, only used a delay of 2 days) and owing to the limited sample sizes used, which restricted the power of the statistical tests employed. Assessing whether the two aspects of structure are comparably effective when a longer retention interval is considered is therefore an important endeavor for future research as it is probably rare for applicants to be interviewed directly after the interviewers have been trained. Furthermore, very little is known about transfer of rating skills to other rating tasks and/or to actual work settings. A study by Noonan and Sulsky (2001) that represents a noteworthy exception in this regard surveyed training participants 4 months after training was provided and found that they reported the use of information taught during the training when completing actual performance evaluations. However, Noonan and Sulsky were not able to assess the actual impact of training on the quality of these real world performance evaluations. Future research is needed to address these gaps in our knowledge.

As another limitation, this study only looked at effects on rating quality but did not take other important criteria into account. Thus, the laboratory

nature of our study precluded the assessment of effects on criterion-related validity. Even though evidence from the AC domain suggests that improvements in rating accuracy after FOR training also lead to better criterion validity (Schleicher et al., 2002), such an effect remains to be shown in the interview domain. Similarly, we did not consider interviewer reactions to the targeted aspects of structure beyond the perception of the difficulty of the rating task. However, previous research has shown that interviewers are not fond of too much structure mainly because structure restricts their freedom in conducting an interview and might thereby interfere with other goals in addition to making valid selection decisions (Dipboye, 1997; Lievens & De Paepe, 2004; van der Zee, Bakker, & Bakker, 2002). Therefore, it might well be that the two components of structure are not comparable when interviewer reactions and thus the adoption of more structure in the long run are taken into account.

We also have to acknowledge that our study tested student participants and that it was conducted in a laboratory setting with videotaped simulated interviews. With regard to the sample employed, past research by Maurer (2002) suggests that the effects of providing descriptive anchors to evaluate interview answers are comparable for students and for job-content experts. Similarly, in the AC domain, Lievens (2001) found comparable effects of rater training when he compared student and managerial samples. These reports provide indirect support for the assumption that the present findings would also be obtained in a field sample with interviewers. With regard to the impact of using videotaped interviews instead of face-to-face interviews, past research has found ambiguous results. No differences were found in a study that compared direct observation of AC candidates to indirect observation (i.e., via video; Ryan et al., 1995). In contrast, however, a study by Van Iddekinge et al. (2006) found higher mean ratings for face-to-face than for videotaped interviews. Furthermore, interrater reliabilities between different interviewers who evaluated the same face-to-face interview were also higher than between two raters of whom one conducted the interview face-to-face whereas the other evaluated the videotaped interviewee. Therefore, future research is needed to assess the external validity of the present findings.

Finally, a reviewer of this paper questioned the content of the training provided to participants in the control condition and raised concerns that this content influenced the results obtained. An alternative to using a pseudotraining condition like in this study would be a control condition that differs only in the presence or absence of FOR training in comparison to the actual training condition. As noted above, the rationale for using the present control condition was to ensure that participants in both training conditions are equally familiar with the training context, the stimulus

materials, and other structural aspects of the training situation. In previous studies, similar pseudotraining control conditions often included a general lecture on performance appraisal instead of relevant components of FOR training (e.g., Roch & O'Sullivan, 2003; Sulsky & Day, 1994). Alternatively, Schleicher et al. (2002) provided more information on ACs in a study in which participants later had to evaluate candidates in different AC exercises. As in these previous studies, we wanted to prevent control participants from providing less accurate ratings just because of their lower familiarity with the rating task or the training situation. Thus, using a no-training control condition would likely have resulted in larger effect sizes for FOR training. However, future research is necessary to determine the actual impact of using different control conditions to evaluate the effectiveness of FOR training.

Concerning additional suggestions for future research, the present findings make it possible that increasing standardization of the evaluation process might be a potential means to reduce subgroup differences and/or discrimination in interviews. Previous research has revealed smaller subgroup differences for higher levels of interview structure (e.g., Huffcutt & Roth, 1998), but again, the coding of interview structure confounded several aspects of structure. Furthermore, in light of the beneficial effects of rater training on AC construct-related validity (Lievens, 2001; Schleicher et al., 2002), future research might also investigate whether increasing interview structure might also help to improve the construct-related validity of dimension ratings generated in selection interviews. Given that past research found little evidence that interviewer ratings actually reflect the targeted dimensions (e.g., Huffcutt et al., 2001; Melchers et al., 2009; Schuler & Funke, 1989), more knowledge on ways to improve interview construct-related validity seems warranted.

Practical Implications

These results show that both FOR rater training for interviewers and descriptively anchored rating scales can lead to substantial improvements in rating accuracy and interrater reliability and that the two methods to increase the structure of the evaluation process in an interview are of comparable effectiveness. Furthermore, the finding that combining them yields further meaningful improvements in rating accuracy is important given that both aim to reduce rater idiosyncrasies and also given the substantial cost of developing good descriptive rating anchors (Maurer, 2002) or of providing rater training for interviewers (Campion et al., 1997). However, inspection of the effect sizes obtained also suggests that it is important to make use of at least one of the two methods to improve structure because this can already help to enhance rating quality considerably. Thus,

when it is not possible to provide both training and anchors, introducing at least one of these two methods can already make a substantial difference. Given that the present results suggest that providing FOR training is of comparable effectiveness to providing descriptively anchored rating scales, the choice between them can be guided by the specific restrictions and possibilities of the context of the interview development and use.

However, we also want to mention two caveats. First, given that we suggest that both methods of improving structure are of comparable effectiveness, it is necessary to keep in mind the limitations of the present study, such as having no substantial time delay between training and interview evaluations. And second, the finding that providing descriptive anchors is as effective as providing rater training should not be interpreted as a suggestion that no training at all is needed for interviewers to attain improvements of rating quality. Training may be needed, for example, to ensure that other methods of interview structure are implemented properly (Campion et al., 1997). For instance, a study by Latham and Saari (1984) revealed that a situational interview that otherwise did possess criterion validity lost its power to predict work performance when used incorrectly by the interviewers (see also Weekley & Gier, 1987, for a similar case).

Furthermore, interviewer training may cover various aspects not included in the current training that focused on the evaluation of answers to predetermined interview questions. Chapman and Zweig's (2005) survey, for example, found that actual interviewer training for campus interviewers often contained information on legal issues, rapport building, or note taking, meaning, on topics beyond those covered by the present training. The importance of such topics can be seen, for instance, in findings from a recent field study by Salimäki and Greenberg (2007), who found that training interviewers in interactional justice can make a marked difference regarding acceptance of job offers, retention of recruited candidates after the probationary period, and organizational commitment of recruited candidates. These findings are in line with the notion that selection interviews serve more purposes than mere selection of the best applicants and that training might also be helpful to achieve those goals. Thus, even though one might want to save money, effort, and time necessary to construct and implement the FOR part of the training, on the basis of these results one should not abandon central topics of interviewer training like the proper use of the interview guide and other important aspects of designing and conducting good selection interviews.

REFERENCES

Athey TR, McIntyre RM. (1987). Effect of rater training on rater accuracy: Levels-of-processing theory and social facilitation theory perspectives. *Journal of Applied Psychology, 72*, 567–572.

Bernardin HJ, Buckley MR, Tyler CL, Wiese DS. (2000). A reconsideration of strategies for rater training. In Ferris GL (Ed.), *Research in personnel and human resources management* (Vol. 18, pp. 221–274). Greenwich, CT: JAI Press.

Bernardin HJ, Pence EC. (1980). Effects of rater training: Creating new response sets and decreasing accuracy. *Journal of Applied Psychology, 65*, 60–66.

Buckley MR, Norris AC, Wiese DS. (2000). A brief history of the selection interview: May the next 100 years be more fruitful. *Journal of Management History, 6*, 113–126.

Burke MJ, Dunlap WP. (2002). Estimating interrater agreement with the average deviation index: A user's guide. *Organizational Research Methods, 5*, 159–172.

Burke MJ, Finkelstein LM, Dusig MS. (1999). On average deviation indices for estimating interrater agreement. *Organizational Research Methods, 2*, 49–68.

Campion MA, Palmer DK, Campion JE. (1997). A review of structure in the selection interview. PERSONNEL PSYCHOLOGY, *50*, 655–702.

Chapman DS, Zweig DI. (2005). Developing a nomological network for interview structure: Antecedents and consequences of the structured selection interview. PERSONNEL PSYCHOLOGY, *58*, 673–702.

Cohen J. (1992). A power primer. *Psychological Bulletin, 112*, 155–159.

Conway JM, Jako RA, Goodman DF. (1995). A meta-analysis of interrater and internal consistency reliability of selection interviews. *Journal of Applied Psychology, 80*, 565–579.

Craik FI. (2002). Levels of processing: Past, present... and future? *Memory, 10*, 305–318.

Craik FI, Lockhart RS. (1972). Levels of processing: A framework for memory research. *Journal of Verbal Learning and Verbal Behavior, 11*, 671–684.

Day DV, Sulsky LM. (1995). Effects of frame-of-reference training and information configuration on memory organization and rating accuracy. *Journal of Applied Psychology, 80*, 158–167.

Deller J, Kleinmann M. (1993). Das situative interview [The situational interview]. In Gebert A, Hacker W (Eds.), *Arbeits- und Organisationspsychologie 1991 in Dresden* (pp. 336–343). Bonn, Germany: Deutscher Psychologen Verlag.

Dipboye RL. (1997). Structured selection interviews: Why do they work? Why are they underutilized? In Anderson N, Herriot P (Eds.), *International handbook of selection and assessment* (pp. 455–473). Chichester, UK: Wiley.

Dipboye RL, Gaugler BB. (1993). Cognitive and behavioral processes in the selection interview. In Schmitt N, Borman WC (Eds.), *Personal selection in organizations* (pp. 135–170). San Francisco, CA: Jossey-Bass.

Dreher GF, Ash RA, Hancock P. (1988). The role of the traditional research design in underestimating the validity of the employment interview. PERSONNEL PSYCHOLOGY, *41*, 315–327.

Dunlap WP, Burke MJ, Smith-Crowe K. (2003). Accurate tests of statistical significance for r_{WG} and average deviation interrater agreement indexes. *Journal of Applied Psychology, 88*, 356–362.

Goodman JS, Wood RE. (2004). Feedback specificity, learning opportunities, and learning. *Journal of Applied Psychology, 89*, 809–821.

Goodman JS, Wood RE. (2009). Faded versus increasing feedback, task variability trajectories, and transfer of training. *Human Performance, 22*, 64–85.

Goodman JS, Wood RE, Hendrickx M. (2004). Feedback specificity, exploration, and learning. *Journal of Applied Psychology, 89*, 248–262.

Highhouse S. (2008). Stubborn reliance on intuition and subjectivity in employee selection. *Industrial and Organizational Psychology: Perspectives on Science and Practice, 1*, 333–342.

Huffcutt AI, Arthur W. (1994). Hunter and Hunter (1984) revisited: Interview validity for entry-level jobs. *Journal of Applied Psychology, 79*, 184–190.

Huffcutt AI, Roth PL. (1998). Racial group differences in employment interview evaluations. *Journal of Applied Psychology, 83*, 179–189.

Huffcutt AI, Weekley JA, Wiesner WH, DeGroot TG, Jones C. (2001). Comparison of situational and behavior description interview questions for higher-level positions. PERSONNEL PSYCHOLOGY, *54*, 619–644.

Huffcutt AI, Woehr DJ. (1999). Further analysis of employment interview validity: A quantitative evaluation of interviewer-related structuring methods. *Journal of Organizational Behavior, 20*, 549–560.

James LR, Demaree RG, Wolf G. (1984). Estimating within-group interrater reliability with and without response bias. *Journal of Applied Psychology, 69*, 85–98.

Klehe U-C, König CJ, Richter GM, Kleinmann M, Melchers KG. (2008). Transparency in structured interviews: Consequences for construct and criterion-related validity. *Human Performance, 21*, 107–137.

Latham GP, Saari LM. (1984). Do people do what they say? Further studies on the situational interview. *Journal of Applied Psychology, 69*, 569–573.

Latham GP, Saari LM, Pursell ED, Campion MA. (1980). The situational interview. *Journal of Applied Psychology, 65*, 422–427.

LeBreton JM, Senter JL. (2008). Answers to 20 questions about interrater reliability and interrater agreement. *Organizational Research Methods, 11*, 815–852.

Lievens F. (2001). Assessor training strategies and their effects on accuracy, interrater reliability, and discriminant validity. *Journal of Applied Psychology, 86*, 255–264.

Lievens F, De Paepe A. (2004). An empirical investigation of interviewer-related factors that discourage the use of high structure interviews. *Journal of Organizational Behavior, 25*, 29–46.

Maas JB. (1965). Patterned scaled expectation interview: Reliability studies on a new technique. *Journal of Applied Psychology, 49*, 431–433.

Macan T. (2009). The employment interview: A review of current studies and directions for future research. *Human Resource Management Review, 19*, 203–218.

Maurer SD. (2002). A practitioner-based analysis of interviewer job expertise and scale format as contextual factors in situational interviews. PERSONNEL PSYCHOLOGY, *55*, 307–327.

Maurer SD, Fay C. (1988). Effect of situational interviews, conventional structured interviews, and training on interview rating agreement: An experimental analysis. PERSONNEL PSYCHOLOGY, *41*, 329–344.

McGraw KO, Wong SP. (1996). Forming inferences about some intraclass correlation coefficients. *Psychological Methods, 1*, 30–46.

Melchers KG, Klehe U-C, Richter GM, Kleinmann M, König CJ, Lievens F. (2009). "I know what you want to know": The impact of interviewees' ability to identify criteria on interview performance and construct-related validity. *Human Performance, 22*, 355–374.

Murphy KR, Cleveland JN. (1995). *Understanding performance appraisal: Social, organizational, and goal-based perspectives*. Thousand Oaks, CA: Sage.

Noonan LE, Sulsky LM. (2001). Impact of frame-of-reference and behavioral observation training on alternative training effectiveness criteria in a Canadian military sample. *Human Performance, 14*, 3–26.

Palmer DK, Campion MA, Green PC. (1999). Interviewing training for both applicant and interviewer. In Eder RW, Harris MM (Eds.), *The employment interview handbook* (pp. 337–351). Thousand Oaks, CA: Sage.

Pulakos ED. (1984). A comparison of rater training programs: Error training and accuracy training. *Journal of Applied Psychology, 69*, 581–588.

Roch SG, O'Sullivan BJ. (2003). Frame of reference rater training issues: Recall, time and behavior observation training. *International Journal of Training and Development, 7*, 93–107.

Ryan AM, Daum D, Bauman T, Grisez M, Mattimore K, Nalodka T, McCormick S. (1995). Direct, indirect, and controlled observation and rating accuracy. *Journal of Applied Psychology, 80*, 664–670.

Ryan AM, McFarland L, Baron H, Page R. (1999). An international look at selection practices: Nation and culture as explanations for variability in practice. PERSONNEL PSYCHOLOGY, *52*, 359–391.

Sackett PR, Lievens F. (2008). Personnel selection. *Annual Review of Psychology, 59*, 419–450.

Salimäki A, Greenberg J. (2007, April). *Attracting applicants and retaining employees by training employment interviewers in interactional justice*. Paper presented at the 22nd Annual Conference of the Society for Industrial and Organizational Psychology, New York.

Salvemini NJ, Reilly RR, Smither JW. (1993). The influence of rater motivation on assimilation effects and accuracy in performance ratings. *Organizational Behavior and Human Decision Processes, 55*, 41–60.

Schleicher DJ, Day DV. (1998). A cognitive evaluation of frame-of-reference rater training: Content and process issues. *Organizational Behavior and Human Decision Processes, 73*, 76–101.

Schleicher DJ, Day DV, Mayes BT, Riggio RE. (2002). A new frame for frame-of-reference training: Enhancing the construct validity of assessment centers. *Journal of Applied Psychology, 87*, 735–746.

Schmidt FL, Zimmerman RD. (2004). A counterintuitive hypothesis about employment interview validity and some supporting evidence. *Journal of Applied Psychology, 89*, 553–561.

Schmidt RA, Bjork RA. (1992). New conceptualizations of practice: Common principles in three paradigms suggest new concepts for training. *Psychological Science, 3*, 207–217.

Schuler H, Funke U. (1989). The interview as a multimodal procedure. In Eder RW, Ferris GR (Eds.), *The employment interview: Theory, research, and practice* (pp. 183–192). Newbury Park, CA: Sage.

Schuler H, Moser K. (1995). Die validität des multimodalen interviews [Validity of the multimodal interview]. *Zeitschrift für Arbeits- und Organisationspsychologie, 39*, 2–12.

Shrout PE, Fleiss JL. (1979). Intraclass correlations: Uses in assessing rater reliability. *Psychological Bulletin, 86*, 420–428.

Stamoulis DT, Hauenstein NM. (1993). Rater training and rating accuracy: Training for dimensional accuracy versus training for ratee differentiation. *Journal of Applied Psychology, 78*, 994–1003.

Sulsky LM, Balzer WK. (1988). Meaning and measurement of performance rating accuracy: Some methodological and theoretical concerns. *Journal of Applied Psychology, 73*, 497–506.

Sulsky LM, Day DV. (1994). Effects of frame-of-reference training on rater accuracy under alternative time delays. *Journal of Applied Psychology, 79*, 535–543.
Sulsky LM, Kline TJB. (2007). Understanding frame-of-reference training success: A social learning theory perspective. *International Journal of Training and Development, 11*, 121–131.
Taylor PJ, Small B. (2002). Asking applicants what they would do versus what they did do: A meta-analytic comparison of situational and past behaviour employment interview questions. *Journal of Occupational and Organizational Psychology, 75*, 277–294.
Uggerslev KL, Sulsky LM. (2008). Using frame-of-reference training to understand the implications of rater idiosyncrasy for rating accuracy. *Journal of Applied Psychology, 93*, 711–719.
van der Zee KI, Bakker AB, Bakker P. (2002). Why are structured interviews so rarely used in personnel selection? *Journal of Applied Psychology, 87*, 176–184.
Van Iddekinge CH, Raymark PH, Roth PL, Payne HS. (2006). Comparing the psychometric characteristics of ratings of face-to-face and videotaped structured interviews. *International Journal of Selection and Assessment, 14*, 347–359.
Van Iddekinge CH, Sager CE, Burnfield JL, Heffner TS. (2006). The variability of criterion-related validity estimates among interviewers and interview panels. *International Journal of Selection and Assessment, 14*, 193–205.
Vance RJ, Kuhnert KW, Farr JL. (1978). Interview judgments: Using external criteria to compare behavioral and graphic scale ratings. *Organizational Behavior and Human Decision Processes, 22*, 279–294.
Weekley JA, Gier JA. (1987). Reliability and validity of the situational interview for a sales position. *Journal of Applied Psychology, 72*, 484–487.
Woehr DJ. (1994). Understanding frame-of-reference training: The impact of training on the recall of performance information. *Journal of Applied Psychology, 79*, 525–534.
Woehr DJ, Huffcutt AI. (1994). Rater training for performance appraisal: A quantitative review. *Journal of Occupational and Organizational Psychology, 67*, 189–205.

APPENDIX

Example of a systematic planning question used in the interview and of the corresponding descriptive rating anchors (the other interview questions are available from the first author upon request):

Tomorrow, you will participate in a full-day advanced training course. You registered for it a while ago but due to the high demand, you had to wait a long time to actually take the course. Two days from now, in the morning, you have to give an important presentation before your board of directors to explain a project you have been planning. Although the basic concept of your presentation is ready, you still need several hours to hone the presentation to perfection in terms of content and style. Later today, however, you have appointments with representatives from other departments. What would you do?

5 (*good*):

If possible, reschedules today's appointments and prepares the presentation so that he/she can attend the training course tomorrow. Would also cancel the training course if necessary, as he/she recognizes the precedence of the presentation over the training course.

3 (*average*):

Reschedules/cancels participation in the training course to have sufficient time to prepare for the presentation.

1 (*poor*):

Accepts a suboptimal presentation (prepared only after the training course) or reschedules the presentation.